A Survey of Statistical Machine Translation

Masao Utiyama

Lecture slides, Kyoto University, August 2006

Contents

- Main components in SMT
- Publicly available SMT system
- Word-based models: IBM models
- Automatic evaluation: BLEU
- Phrase-based SMT

Contents (cont.)

- Decoding
- Log-linear model
- Minimum error rate training
- Reranking
- Discriminative SMT

Contents (cont.)

- Formally syntax-based SMT
- Formally and linguistically syntax-based SMT
- Dependency-based SMT
- Word alignment
- Parallel corpora
- Next topics in SMT (in my opinion)

Main components in SMT

- Parallel corpora
- SMT engine
- Automatic evaluation

Publicly available SMT system

http://www.statmt.org/wmt06/shared-task/

Philipp Koehn and Christof Monz. (2006) Manual and Automatic Evaluation of Machine Translation between European Languages. HLT-NAACL workshop on Statistical Machine Translation.

- parallel corpora
- word alignment program
- decorder
- minimum error rate training package

Performance of baseline system

- speed: a few seconds per sentence
- accuracy: comparable to SYSTRAN?
- languages: any language pair with a large parallel corpus

  Accuracy depends on the language pairs.

IBM models

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. (1993) The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, 19:2, 263–311.

- Translate a French sentence $\mathbf{f}$ into an English sentence $\mathbf{e}$

$$\Pr(\mathbf{e}|\mathbf{f}) = \frac{\Pr(\mathbf{e})\Pr(\mathbf{f}|\mathbf{e})}{\Pr(\mathbf{f})}$$

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{e})\Pr(\mathbf{f}|\mathbf{e})$$

Three components in SMT

$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{f}|\mathbf{e})$

- Language model: $\Pr(\mathbf{e})$

  Responsible for fluency

- Translation model: $\Pr(\mathbf{f}|\mathbf{e})$

  Responsible for adequacy

- Search or decoding: $\arg\max_{\mathbf{e}}$

  Responsible for candidate selection

Word alignment $\mathbf{a}$

- Le$_1$ programme$_2$ a$_3$ été$_4$ mis$_5$ en$_6$ application$_7$
- And(0) the(1) program(2) has(3) been(4) implemented(5,6,7)
- J'$_1$ applaudis$_2$ á$_3$ la$_4$ décision$_5$
- e$_0$(3) I(1) applaud(2) the(4) decision(5)

Translation model as sum of alignments

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

- $\mathbf{e} = e_1^l = e_1 e_2 \ldots e_l$ (e_i is i-th English word)

- $\mathbf{f} = f_1^m = f_1 f_2 \ldots f_m$ (f_j is j-th French word)

- $\mathbf{a}_1^m = a_1 a_2 \ldots a_m$ ($a_j = i$ if j-th French position is connected to i-th English position. $a_j = 0$ if f_j is not connected to any English word)

Sequence model

$$
\begin{aligned}
\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}) &= \text{(Choose the length of } \mathbf{f}\text{)} \\
&\quad \times \prod_{j=1}^{m} \text{(Choose where to connect } j\text{-th French word)} \\
&\qquad\qquad \times \text{(Choose } j\text{-th French word)} \\
&= \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j | a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j | a_1^{j}, f_1^{j-1}, m, \mathbf{e})
\end{aligned}
$$

Model 1

- $\Pr(m|\mathbf{e}) = \epsilon$ (a constant independent of m and $\mathbf{e}$)
- $\Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) = \frac{1}{l+1}$ (depends only on l, the length of $\mathbf{e}$)
- $\Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}) = t(f_j|e_{a_j})$ (translation probability of f_j given e_{a_j})

$t(f_j|e_{a_j})$ is estimated by EM.

HMM-based word alignment

Stephan Vogel, Hermann Ney, and Cristoph Tillmann. (1996) HMM-Based Word Alignment in Statistical Translation. COLING-96.

- $\Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) = \Pr(a_j|a_{j-1}, l)$ (depends only on the previous alignment a_{j-1} and the length of $\mathbf{e}$)
- Probability depends on the relative position

$$\Pr(a_j|a_{j-1}, l) = \frac{s(a_j - a_{j-1})}{\sum_{i=1}^{l} s(i - a_{j-1})}$$

$\{s(\cdot)\}$ is a set of non-negative parameters. The empty word is not considered.

Fertility model

$\phi_i =$ fertility of i-th English word (no. of French words to which e_i is connected)

$\tau_{ik} = k$-th $(1 \leq k \leq \phi_i)$ French words to which e_i is connected

$\pi_{ik} =$ position in $\mathbf{f}$ of τ_{ik}

$$\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}) = \sum_{(\tau,\pi) \in \langle \mathbf{f}, \mathbf{a} \rangle} \Pr(\tau, \pi | \mathbf{e})$$

- Given an English sentence, $\mathbf{e}$
- Decide the fertility of each English word,
- Choose a list of French words (τ) to connect to each English word,
- Choose a position of each French word (π)

$$\begin{aligned}
\Pr(\tau, \pi) &= \prod_{i=1}^{l} \Pr(\phi_i | \phi_1^{i-1}, \mathbf{e}) \text{(Choose fertility of each non-empty word } e_i\text{)} \\
&\times \Pr(\phi_0 | \phi_1^{l}, \mathbf{e}) \text{(Choose fertility of the empty word } e_0\text{)} \\
&\times \prod_{i=0}^{l} \prod_{k=1}^{\phi_i} \Pr(\tau_{ik} | \tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^{l}, \mathbf{e}) \text{(Choose } k\text{-th French word for } e_i\text{)} \\
&\times \prod_{i=1}^{l} \prod_{k=1}^{\phi_i} \Pr(\pi_{ik} | \pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^{l}, \phi_0^{l}, \mathbf{e}) \\
&\quad \text{(position of each French word } \tau_{ik} \text{ connecting to non-empty word } e_i\text{)} \\
&\times \prod_{k=1}^{\phi_0} \Pr(\pi_{0k} | \pi_{01}^{k-1}, \pi_1^{l}, \tau_0^{l}, \phi_0^{l}, \mathbf{e}) \\
&\quad \text{(position of each French word connecting to the empty word)}
\end{aligned}$$

Model 3

- $n(\phi|e_i) = \Pr(\phi_i|\phi_1^{i-1}, \mathbf{e})$ (fertility probability for non-empty words)

- $t(f|e_i) = \Pr(T_{ik} = f|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e})$ (translation probability for all words)

- $d(j|i, m, l) = \Pr(\Pi_{ik} = j|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e})$ (distortion probability for non-empty words)

- $\prod_{k=1}^{\phi_0} \Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}) = \frac{1}{\phi_0!} = \frac{1}{\phi_0(\phi_0-1)(\phi_0-2)\cdots}$. ($\tau_{01}$ has ϕ_0 positions to go. τ_{02} has $\phi_0 - 1$ positions, and so on)

- $\Pr(\phi_0|\phi_1^l, \mathbf{e}) = \begin{pmatrix} \phi_1 + \phi_2 + \cdots + \phi_l \\ \phi_0 \end{pmatrix} p_0^{\phi_1+\phi_2+\cdots+\phi_l-\phi_0} p_1^{\phi_0}$

IBM models summary

Automatic construction of a machine translation system is feasible given a parallel corpus.

Automatic evaluation: BLEU

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. ACL-02.

The closer a machine translation is to a professional human translation, the better it is.

Essential ingredients:

- a numerical "translation closeness" metric
- a corpus of good quality human reference translations

BLEU

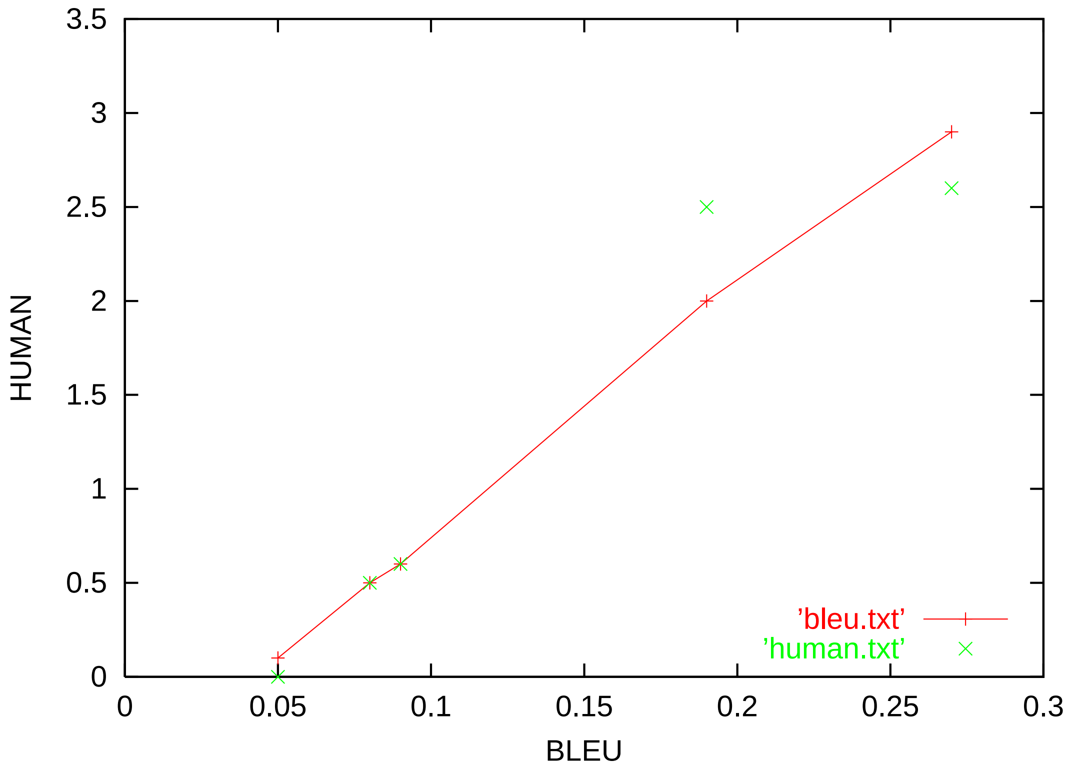
$$\text{BLEU} = \text{BP} \times \exp(\sum_{n=1}^{N} \frac{1}{N} \log p_n)$$

- BP: brevity penalty (punish short translations)
- $p_n = \text{ngram precision} = \frac{\text{clipped number of shared n-grams}}{\text{number of ngrams in translations}}$

Evaluating BLEU

- 5 Chinese-to-English MT (S1,S2,S3,H1,H2) were evaluated
- 250 Chinese sentences were translated
- monolingual and bilingual groups evaluated translations

Relationship between BLEU and human evaluations



BLEU summary

- BLEU correlates highly with human evaluation
- BLEU is used in all SMT papers these days

BLEU limitations

Chris Callison-Burch, Miles Osborne and Philipp Koehn. (2006) Re-evaluating the Role of BLEU in Machine Translation Research. EACL-06.

appropriate use of BLEU

- Comparing systems with similar translation strategy
- Parameter optimization

inappropriate use of BLEU

- Comparing systems with different translation strategy (SMT vs rule-based MT)
- Trying to detect improvements not modelled well by BLEU

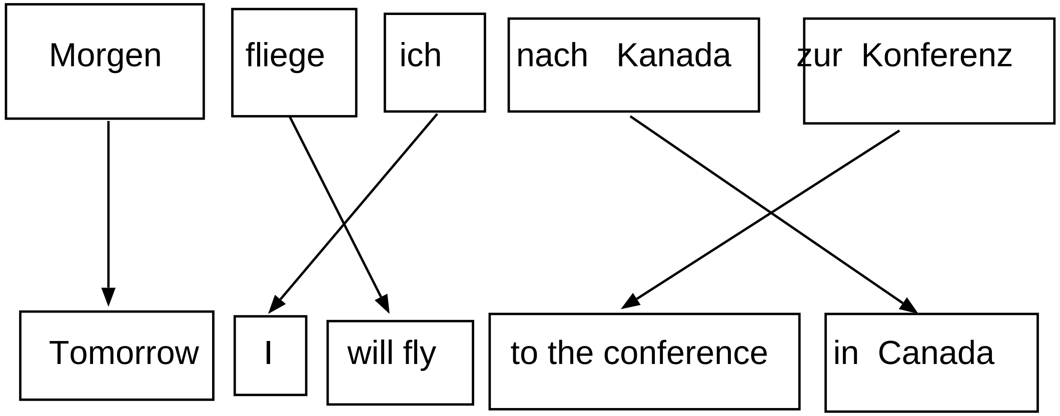
Phrase-based models

Franz Josef Och and Hermann Ney. (2004) The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30:4, 417–449.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. (2003) Statistical Phrase-Based Translation. HLT-NAACL-2003

$Phrase$ refers to a consecutive sequence of words occurring in text.

Phrase-based translation



Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

from Koehn et al. 2005.

Phrase-based model by Koehn, Och and Marcu 2003

- $\mathbf{f}$ is segmented into a sequence of I phrases $\overline{f_1^I}$
- Each $\overline{f_i}$ is translated into an English phrase $\overline{e_i}$
- Phrase translation probabilty is $\phi(\overline{f_i}|\overline{e_i})$
- Reordering of the English output phrases is modeled by a relative distortion probability distribution $d(a_i - b_{i-1})$
- a_i is the start position of $\overline{f_i}$
- b_{i-1} is the end position of $\overline{f_{i-1}}$

Phrase-based model (cont.)

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \Pr(\mathbf{e}) \Pr(\mathbf{f}|\mathbf{e}) \omega^{\mathsf{length(e)}}$$

$$\Pr(\mathbf{f}|\mathbf{e}) = \prod_{i}^{I} \phi(\overline{f_i}|\overline{e_i}) d(a_i - b_{i-1})$$

- $d(a_i - b_{i-1})$ is trained using a joint probablity model
- $d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1}|}$ is used in the Pharaoh decoder

Research issues in phrase-based models

- How to extract phrases?
- How to estimate $\phi(\overline{f}|\overline{e})$?

  $\phi(\overline{f}|\overline{e}) = \frac{\mathsf{count}(\overline{f}|\overline{e})}{\sum_{\overline{f}} \mathsf{count}(\overline{f}|\overline{e})}$
- How to estimate the distortion or reordering probability distribution $d(\cdot)$?

  Kohen et al.'s model is very simple. Much recent work focuses on reordering.

Phrase extraction

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne and David Talbot. (2005) Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. IWSLT-2005.

- Align the words bidirectionally by GIZA++
- Intersection: high precision alignment
- Union: high recall alignment
- Extract phrase paris that are consistent with the word alignment

Phrase extraction

	Maria	no	daba	una	botefada	a	la	bruja	verde
Mary	■								
did		▨				▨			
not		■							
slap			▨	▨	■				
the							■		
green									■
witch								■	

Mary	Maria
did	no
did not	no
Mary did not	Maria no
slap	daba
slap	una
slap	botefada
slap	daba una
slap	daba una botefada
...	

Phrase-based SMT summary

- Currently, best performing SMT.
- A phrase translation pair incorporates local reordering
- Heuristic phrase extraction works well
- Global reordering is a research issue

Phrase-based models are constructed on top of the word-based IBM models.

Theoretical issues in phrase-based models

- Phrase-based models are constructed on top of the word-based IBM models.
- A generative phrase-based model is inferior to heuristic models.

Daniel Marcu and William Wong. (2002) A phrase-based, joint probability model for statistical machine translation. EMNLP-02.

Decoding or Search

Different models require different decoders. Basic algorithms are borrowed from speech recognition or parsing.

- Beam search

  Franz Josef Och and Hermann Ney (2004)

- Weighted finite state transducer

  Kevin Knight and Yaser Al-Onaizan. (1998) Translation with Finite-State Devices. 4th AMTA

- CYK

  Syntax-based SMT

Log-linear model

Franz Josef Och. (2003) Minimum Error Rate Training in Statistical Machine Translation. ACL-2003.

$$\begin{aligned}
\hat{\mathbf{e}} &= \arg\max_{\mathbf{e}} \Pr(\mathbf{e}|\mathbf{f}) \\
&= \arg\max_{\mathbf{e}} \frac{\exp(\sum_m \lambda_m h_m(\mathbf{e},\mathbf{f}))}{\sum_{\mathbf{e}'} \exp(\sum_m \lambda_m h_m(\mathbf{e}',\mathbf{f}))} \\
&= \arg\max_{\mathbf{e}} \exp(\sum_m \lambda_m h_m(\mathbf{e},\mathbf{f}))
\end{aligned}$$

- λ_m = weight of feature m
- $h_m(\mathbf{e},\mathbf{f})$ = value of feature m

Research issues in log-linear models

- Modeling problem

  Developing suitable feature functions that capture the relevant properties of the translation task

- Training problem

  obtaining suitable feature weights

Example of features

- language model probability: $\Pr(e)$
- phrase translation probability: $\phi(\overline{e}|\overline{f})$, $\phi(\overline{f}|\overline{e})$
- lexical translation probability: $t(e_i|f_j)$, $t(f_i|e_j)$
- word and phrase penalty (prefer long or short sentences)

Minimum error rate training

Optimize the weights of features λ

1. running the decorder with a currently best weight setting
2. extracting an n-bset list of possible translations
3. finding a better weight setting that re-ranks the n-best-list, so that a better translation score is obtained.
4. goto 1

Minimum Bayes-risk decoding

Ashish Venugopal, Andreas Zollmann and Alex Waibel. (2005) Training and Evaluating Error Minimization Rules for Statistical Machine Translation. ACL Workshop on Building and Using Parallel Texts.

$$\hat{\mathbf{e}} = \arg \min_{\mathbf{e} \in \mathsf{Gen}(\mathbf{f})} \sum_{\mathbf{e}' \in \mathsf{Gen}(\mathbf{f})} \mathsf{Loss}(\mathbf{e}, \mathbf{e}') \Pr(\mathbf{e}' | \mathbf{f})$$

$\hat{\mathbf{e}}$ is regarded as the centroid of Gen($\mathbf{f}$) if we regard Loss($\mathbf{e}$,$\mathbf{e}'$) as the distance between $\mathbf{e}$ and $\mathbf{e}'$. In other word, $\hat{\mathbf{e}}$ is the representative translation of Gen($\mathbf{f}$).

Comparison of performance

Minimum Bayes risk (MBR) $>$(slighly) MAP with Minimum error training

1. MAP returns the top ranked translation
2. MBR examines the n-best list for the best translation

N-best reranking

Franz Josef Och, Daniel Glidea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. (2004) A Smorgasbord of Features for Statistical Machine Translation. HLT-NAACL-2004.

Libin Shen, Anoop Sarkar, and Franz Josef Och. (2004) Discriminative Reranking for Machine Translation. HLT-NAACL-2004.

Reranks the n-best list of translation candidates.

N-best reranking (cont.)

- Easy to incorporate various features including global features that are difficult to handle in decoding

- Currently, improvement with reranking is modest

- This is because the translations contained in an n-best list are quite limited compared to the vast search space

Log-linear model summary

- Log-linear model is one of the best practices in the current SMT

Discriminative SMT

Percy Liang, Alexandre Bouchard-Cote, Dan Klein, and Ben Taskar. (2006) An End-to-End Discriminative Approach to Machine Translation. ACL-2006.

- Formuate SMT as structured classification
- Represent a translation candidate as a feature vector
- Estimate a weight vector from a large training corpus, in contrast to minimum error training that uses a small development corpus

Translation as structured classification

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e},\mathbf{h}} \mathbf{w} \cdot \Phi(\mathbf{f}, \mathbf{e}, \mathbf{h})$$

- $\hat{\mathbf{e}}$ is the English translation
- $\mathbf{f}$ is the input foreign language sentence
- $\mathbf{e}$ is a candidate translation
- $\mathbf{h}$ is a hidden structure
- $\mathbf{w}$ is a weight vector
- $\Phi(\mathbf{f}, \mathbf{e}, \mathbf{h})$ is a feature vector representing $(\mathbf{f}, \mathbf{e}, \mathbf{h})$

$$\mathbf{w} \cdot \Phi(\mathbf{f}, \mathbf{e}, \mathbf{h})$$

- A candidate $\mathbf{e}$ is represented by a feature vector $\Phi(\mathbf{f}, \mathbf{e}, \mathbf{h})$
- Its score is $\mathbf{w} \cdot \Phi(\mathbf{f}, \mathbf{e}, \mathbf{h})$
- Score calculation and candidate generation is done by a decoder.
- $\mathbf{w}$ is updated for each $(\mathbf{f}_i, \mathbf{e}_i)$

Once you can represent a translation candidate as a feature vector, you can tune a weight vector to search better translations.

Perceptron-based training

Update rule on an example $(\mathbf{f}_i, \mathbf{e}_i)$

$$\mathbf{w} \leftarrow \mathbf{w} + \Phi(\mathbf{f}_i, \mathbf{e}_t, \mathbf{h}_t) - \Phi(\mathbf{f}_i, \mathbf{e}_p, \mathbf{h}_p)$$

- $\mathbf{e}_p, \mathbf{h}_p$ are the predicted translation and hidden structure, i.e., the best translation obtained by a decorder.
- $\mathbf{e}_t, \mathbf{h}_t$ are the target translation and hidden structure.
- To obtain $\mathbf{e}_t$, generate an n-bset list using the current parameters. Then select the highest BLEU score translation in the n-best list wrt $\mathbf{e}_i$

Discriminative SMT summary

- A large number of feature weights can be tuned
- Use all training data for tuning weights

Reordering for phrase-based SMT

Kazuteru Ohashi, Kazuhide Yamamoto, Kuniko Saito, and Masaaki Nagata. (2005) NUT-NTT Statistical Machine Translation System for IWSLT 2005.

$$\Pr(\mathbf{f}|\mathbf{e}) = \prod_{i}^{I} \phi(\overline{f_i}|\overline{e_i}) d(a_i - b_{i-1})$$

- Reordering in phrase-based SMT is very weak.
- Simply to penalize nonmonotonic phrase alignment
- Can't represent the general tendency of global phrase reordering

Phrase distortion model

$$p(d|\overline{e}_{i-1}, \overline{e}_i, \overline{f}_{i-1}, \overline{f}_i)$$

- $\overline{e}_{i-1}, \overline{e}_i$ are adjacent target phrases
- $\overline{f}_{i-1}, \overline{f}_i$ are source phrases aligned
- d is the relative distance between f_{i-1} - f_i

Slightly improvements in accuracy.

Discriminative reordering model

Richard Zens and Hermann Ney. (2006) Discriminative Reordering Models for Statistical Machine Translation. HLT-NAACL-2006 Workshop on Statistical Machine Translation.

$$p(o|\overline{e}_{i-1}, \overline{e}_i, \overline{f}_{i-1}, \overline{f}_i)$$

- $o =$ orientation (left/right)
- o is left if the start position of $\overline{f}_i$ is to the left of $\overline{f}_{i-1}$

Training

- Maximum entropy modeling
- Word alignment corpora were used for training corpora
- Words and word classes were used as features

Formally syntax-based SMT

Dekai Wu. (1997) Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. Computational Linguistics, 23:3, 377–403.

David Chiang. (2005) A Hierarchical Phrase-Based Model for Statistical Machine Translation. ACL-2005.

Learn a synchronous context-free grammer from a bitext without any syntactic information.

Hierarchical phrases

［オーストラリア］ ［是］ ［［［与 ［北朝鮮］1 有 ［外交］2］ 的 ［少数国家］3］ 之一］

［Australia］ ［is］ ［one of ［the ［few countries］3 that ［have ［diplomatic relations］2 with ［North Korea］1］］］

- 是=is, 与=with, 有=have, 的=that, 之一=one of
- 与 X1 有 X2, have X2 with X1
- X1 = 北朝鮮, North Korea
- X2 = 外交, diplomatic relations

Hierarchical phrases (cont.)

［オーストラリア］ ［是］ ［［［与 ［北朝鮮］1 有 ［外交］2］ 的 ［少数国家］3］ 之一］

［Australia］ ［is］ ［one of ［the ［few countries］3 that ［have ［diplomatic relations］2 with ［North Korea］1］］］

- X1 的 X2, the X2 that X1
- X1 = 与 北朝鮮 有 外交, have diplomatic relations
- X2 = 少数国家, few countries

Hierarchical phrases (cont.)

［オーストラリア］［是］［[［与［北朝鲜］1 有［外交］2］的［少数国家］3］之一］

［Australia］［is］［one of［the［few countries］3 that［have［diplomatic relations］2 with［North Korea］1]]]

- X1 之一, one of X1
- X1 = 与 北朝鲜 有 外交 的 少数国家, the few countries that have diplomatic relations with North Korea

Parsing = translation

```
オーストラリア -------------------------------------------(Australia)X--S-+
是 --------------------------------------------------------------(is)X----+-S---+
与 -----------------+                                                           |
北朝鮮 -(N. Korea)X1+                                                           |
有 -----------------+-(have X2 with X1)X1-+                                     |
外交 -(dipl. rel.)X2+                     |                                     |
的 ---------------------------------------+--(the X2 that X1)X1--+              |
少数国家 ----(few countries)X2 -----------+                      |              |
之一 ------------------------------------------------------------+-- one of X1 -+- S
```

Rule acquisition

1. If $< \overline{f}, \overline{e} >$ is an initial phrase pair, then

   $X \rightarrow < \overline{f}, \overline{e} >$

   is a rule.

2. If $r = X \rightarrow < \gamma, \alpha >$ is a rule and $< \overline{f}, \overline{e} >$ is an initial phrase pair such that $\gamma = \gamma_1 \overline{f} \gamma_2$ and $\alpha = \alpha_1 \overline{e} \alpha_2$ then

   $X \rightarrow < \gamma_1 X_1 \gamma_2, \alpha_1 X_1 \alpha_2 >$

   is a rule. (Replace initial phrases with variables)

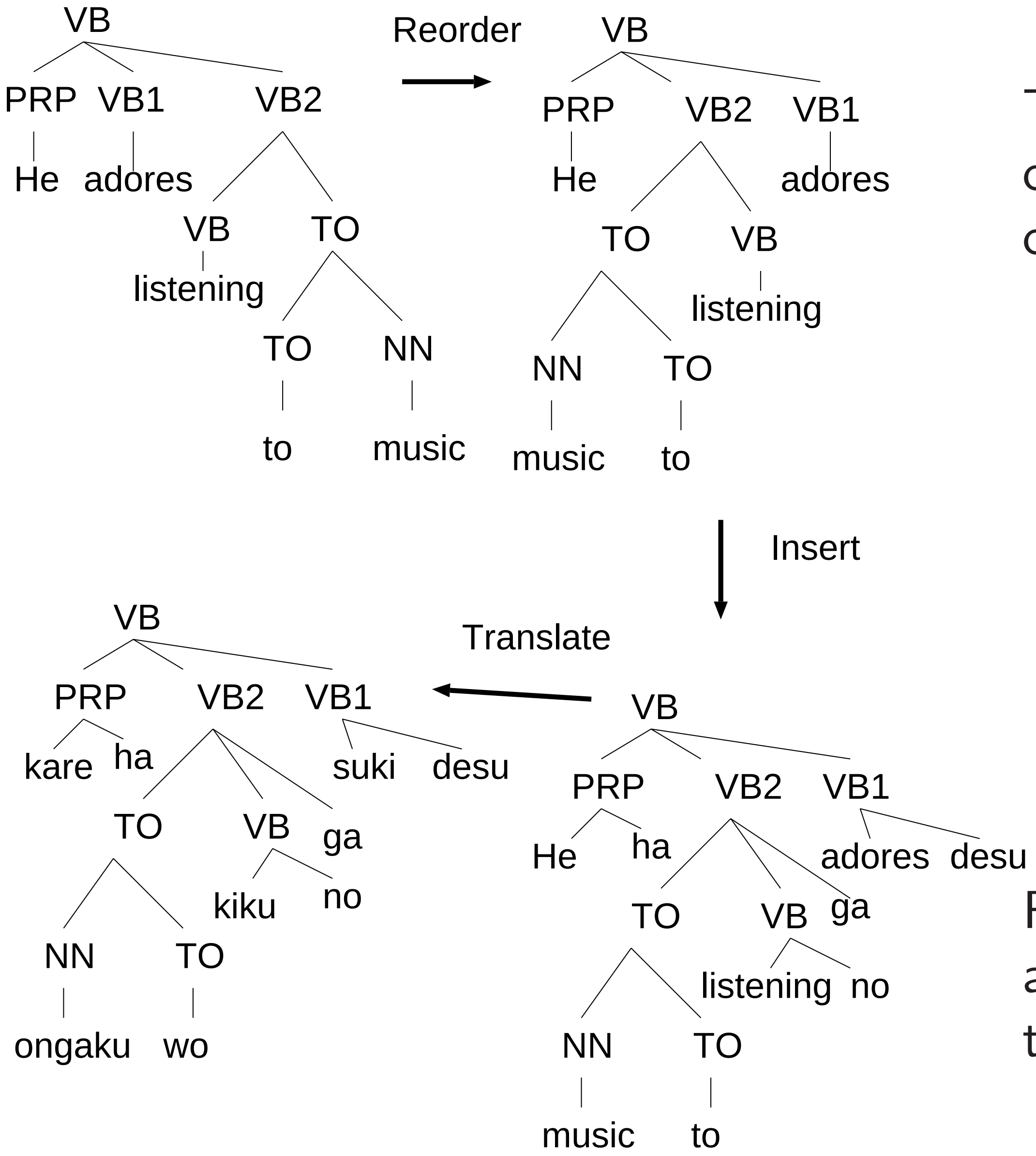
Linguistically and formally syntax-based SMT

Kenji Yamada and Kevin Knight. (2001) A Syntax-Based Statistical Translation Model. ACL-2001.

Kenji Yamada and Kevin Knight. (2002) A Decoder for Syntax-based Statistical MT. ACL-2002.

Parse a foreign sentence as an English sentence when translating $\mathbf{f}$ to $\mathbf{e}$.

- training corpus = parsed English sentences and raw foreign sentences



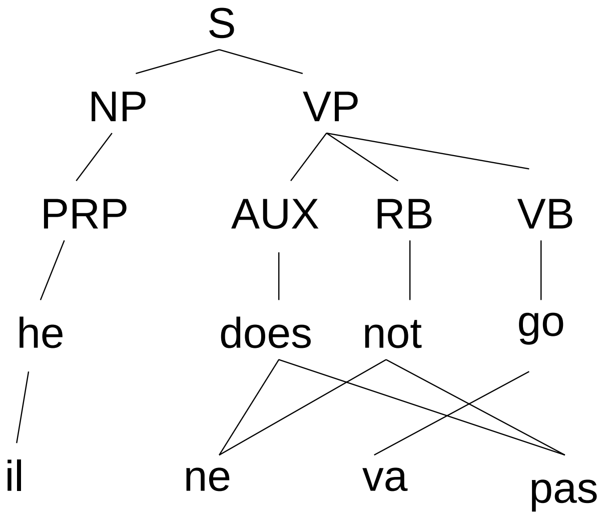
Three kinds of operations are applied on each node:

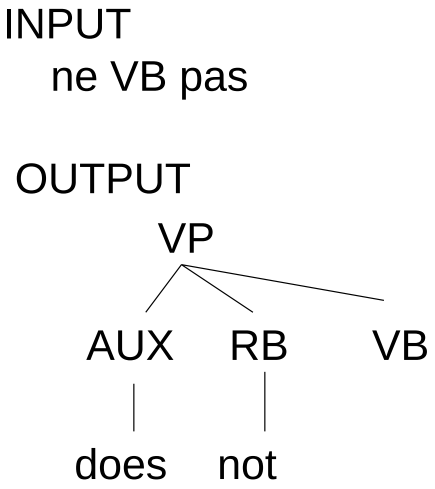
- reordering child nodes
- inserting an extra word to the left or right of the node
- translating leaf words

Reverse operations are performed when translating $\mathbf{f}$ into $\mathbf{e}$

Rule acquisition

Michel Galley, Mark Hopkins, Kevin Knight and Daniel Marcu. (2004) What's in a Translation Rule? NAACL-HLT-2004.




INPUT
ne VB pas

OUTPUT

Transformation rules are extracted from alignment graphs

Dependency-based SMT

Chris Quirk, Arul Menezes, and Colin Cherry. (2005) Dependency Treelet Translation: Syntactically Informed Phrasal SMT. ACL-2005.

Yuan Ding and Martha Palmer. (2005) Machine Translation Using Probabilistic Synchronous Dependency Insertion Grammars. ACL-2005.

David A. Smith and Jason Eisner. (2006) Quasi-Synchronous Grammars: Alignment by SOft Projection of Syntactic Dependencies. HLT-NAACL-2006 Workshop on Statistical Machine Translation.

Dependency Treelet Translation

1. Align a parallel corpus

2. Project the source dependency parse onto the target sentence

3. Extract dependency treelet translation pairs

4. Train a tree-based ordering model

Projecting dependency trees

1. Dependency projection (English dependency -> French dependency)

```
            +-[0]startup    de demarrage[0]-+
    +-[1]properties                    proprietes[1]-+
[2]and                                             et[2]
    +-[3]options                          options[3]-+
```

2. Reattachment (to correct the word order in the French sentence)

```
            +-[0]startup    proprietes[1]--------+
    +-[1]properties                           et[2]
[2]and                              options[3] --+
    +-[3]options                de demarrage[0] -+
```

Translation model

$$\prod_{t,s} \Pr(\text{order}(t)|t,s)\,\Pr(t|s)$$

- Treelet s collectively covers the source dependency tree
- t is the translation of s
- order(t) is the order of the constituent of t
- order(t) is learned by a decision tree

Advantage of dependency translation

- Discontiguous phrases
- Reordering based on syntax
- Dependency is better suited to lexicalized models compared with constituency analysis

Phrase cohesion

Heidi J. Fox. (2002) Phrasal Cohesion and Statistical Machine Translation. EMNLP-2002.

Explore how well phrases cohere across two languages, specifically English and French.

crossing

- $\mathbf{e}_i^j = e_i \ldots e_j$
- span$(\mathbf{e}_j^i) = f_{min(a_i,\ldots,a_j)}, \ldots, f_{max(a_i,\ldots,a_j)}$

If phrase cohere perfectly across languages, the span of one phrase will never overlap the span of another. If two spans do overlap, we call this a crossing.

Phrase cohesion (cont.)

Head-modifier crossing

	#crossing per sentence
constituency analysis	2.252
dependency analysis	1.88

Modifier-modifier crossing

	#crossing per sentence
constituency analysis	0.86
dependency analysis	1.498

Reordering words by phrasal movement is a reasonable strategy

The highest degree of cohesion is present in dependency structures

Word alignment

Franz Josef Och and Hermann Ney. (2003) A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29:1, 19–51.

Robert C. Moore (2005) A Discriminative Framework for Bilingual Word Alignment. HLT-NAACL-2005.

Systematic comparison of alignment models

- Statistical alignment models outperform the simple Dice coefficient
- The best results are obtained with Model-6 (log-linear interpolation of Model-4 and HMM-model)
- Smoothing and symmetrization have a significant effect on the alignment quality

Discriminative word alignment

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \sum_i \lambda_i f_i(\mathbf{a}, \mathbf{e}, \mathbf{f})$$

features

- Association score
- Nonmonotonicity features
- one-to-many feature
- unlinked word feature

Parameter optimization

A modified version of averaged perceptron learning

$$\lambda_i \leftarrow \lambda_i + \eta(f_i(\mathbf{a}_{\text{ref}}, \mathbf{e}, \mathbf{f}) - f_i(\mathbf{a}_{\text{hypo}}, \mathbf{e}, \mathbf{f}))$$

- $\mathbf{a}_{\text{ref}}$ = reference alignment between $\mathbf{e}$ and $\mathbf{f}$
- $\mathbf{a}_{\text{hypo}}$ = best alingment obtained by a beam search

Comparable performance with IBM Model 4.

Language model

Eugene Charniak, Kevin Knight and Kenji Yamada. Syntax-based Language Models for Statistical Machine Translation

Syntax-based language models improve the syntactic quality of SMT.

Almut Silja Hildebrand, Matthias Eck, Stephan Vogel and Alex Waibel. (2005) Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. EAMT 2005.

Select part of a training parallel corpus that are similar to the test sentences.

Bing Zhao, Mathias Eck and Stephan Vogel. Language Model Adaptation for Statistidcal Machine Translation with Structured Query Models.

Select part of a target monolingual corpus that are searched by CLIR techniques.

Construction of parallel corpora

Masao Utiyama and Hitoshi Isahara. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. ACL-2003

Dragos Stefan Munteanu and Daniel Marcu. (2006) Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Computational Linguistics 31:4, 477–504.

Matthias Eck, Stephan Vogel and Alex Waibel. (2005) Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF. IWSLT-2005.

Other topics

Evgeny Matusov, Nicota Ueffing and Hermann Ney. (2006) Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. EACL-2006.

Chris Callison-Burch, Philipp Koehn and Miles Osborne. (2006) Improved Statistical Machine Translation Using Paraphrases.

Franz Josef Och and Hermann Ney. What Can Machine Translation Learn from Speech Recognition?

Next topics in SMT (in my opinion)

- SMT from comparable corpora
- Reliable human evaluation
- Automatic error analysis
- Language modeling
- SMT between non-English languages
- Minimum error rate phrase extraction

SMT from comparable corpora

- Current phrase-based translation models are based on phrase tables.
- Thus, if we can construct phrase tables from comparable corpora, we can make a statistical machine translation system from comparable corpora.
- We can assume a small development parallel corpora for tuning parameters.

cf. (Dragos Stefan Munteanu and Daniel Marcu 2006)

Reliable human evaluation

- Judgments on adequacy and fluency are unstable (Koehn and Monz 2006)
- More reliable human judgments are needed.
- Candidates:
  - paired comparison analysis
  - ranking

Automatic error analysis

- BLEU is useful for ranking systems
- However, we want to spot the type of errors
- Better evaluation is needed.
- The first step is large scale (automatic) evaluation
- Then, we can classify test sentences into good and bad translations
- We can extract characteristic error patterns in bad translations

Identifying what can do and what can not do.

Language modeling

- Very large scale n-gram model (20 billion?)
- Maximum entropy modeling
- Discriminative language modeling
- Rich features

SMT between non-English languages

Parallel corpus

- There are a large amount of English-Chinese parallel corpora
- There are a large amount of English-Arabic parallel corpora
- However, Chinese-Arabic parallel corpora are few if exist at all

Pivot translation is promising for the time being.

SMT between non-English languages (cont.)

- Morphology
- Reordering

A language universal approach is needed which can be applied to all language pairs.

Minimum error rate phrase extraction

- Phrases have large impact on MT performance
- Connecting phrase extraction and decoding is needed.

Cf. (Nagata et. al 2006)

Difficult problems

- Better modeling
- Better evaluation
- Publicly available large scale multilingual corpora

ACL-2006

Incomplete notes taken while ACL-2006

Arabic-to-English SMT

Fatiha Sadat and Nizar Habash. Combination of Arabic Preprocessing Schemes for Statistical Machine. Translation.

Arabic Linguistic Issues

- Rich morphology
- Spelling ambiguity
- Segmentation Ambiguity

Preprocessing can improve MT quality.

Word alignment error

Necip Fazil Ayan and Bonnie J. Dorr. Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT

- AER and BLEU are not correlated very much
- Precision-oriented alignments is better than recall-oriented alignments in phrase-based SMT

Word alignment with CRF

Phil Blunsom and Trevor Cohn. Discriminative Word Alignment with Conditional Random Fields.

- Conditional random fields based on the HMM alignment model
- HMM alignment model is a sequential model
- Alignment as tagging.
- For each source word, assign a position in the target sentence

Phrase extraction from comparable corpora

Dragos Stefan Munteanu and Daniel Marcu. Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora.

- Begin with a word alignment sentence pair
- Detect fragments (phrases) that are translation with each other
- Alignment score of a fragment is the average of word alignment scores in a window

Word alignment

Robert C. Moore, Wen-tau Yih and Andreas Bode. Improved Discriminative Bilingual Word Alignment

Two stage approach: Stage 1 aligns words then Stage 2 uses the results of stage 1 to align words. Stage1 features are association score, jump distance differences, etc. Stage2 features are conditional link cluster odds, ...

New features had a large impact on performance

perceptron learning $<$ Joachims' SVM^{struct} (note!)

Reordering

Deyi Xiong, Qun Liu and Shouxun Lin. Maximum Entropy Based Phrase Reordering Model for Statistical Machine Translation

$$\Pr(o|P_1, P_2)$$

P_1 and P_2 are phrases. o is either straight or inverted.

- Translation model is built upon BTG.
- CYK-style decoder

Reordering

Distortion Models for Statistical Machine Translation Yaser Al-Onaizan and Kishore Papineni

Reordering scrambled English

- Input: English written in foreign language order
- Task: Recover English order.
- N-gram $(3 \leq N \leq 5)$ language model is not sufficient to address word order issue in SMT

Distortion model

$\Pr($distance between f_i and $f_j | f_i, f_j)$

Syntax-based SMT

Tree-to-String Alignment Template for Statistical Machine Translation Yang Liu, Qun Liu and Shouxun Lin

Tree-string-alignment (TSA): bilingual phrase with tree over the source string

Corporation between phrases (n-grams) and syntactic constituents are not easy.

Reordering

Masaaki Nagata, Kuniko Saito, Kazuhide Yamamoto and Kazuteru Ohashi. A Clustered Global Phrase Reordering Model for Statistical Machine Translation

$p(d|\overline{e}_{i-1}, \overline{e}_i, \overline{f}_{i-1}, \overline{f}_i)$

- $d =$ monotone ajacent, monotone gap, reverse adjacent, reverse gap
- mkcls was used to cluster phrases
- Jointly segmenting and aligning phrases performs better than grow-diag-final. (note!)

Discriminative SMT

Christoph Tillmann and Tong Zhang. A Discriminative Global Training Algorithm for Statistical MT.

Block bigram model: $(b_1, N,_{)}, (b_2, L, b_1), ...$

feature: $f(b_i, o, b_{i-1})$

$s_w(b_1^n, o_1^n) = \sum_{i=1}^n w^T \cdot f(b_i, o_i, b_{i-1})$

$\arg\max s_w(b_1^n, o_1^n)$

Discriminative SMT

Percy Liang, Alexandre Bouchard-Cote, Dan Klein and Ben Taskar. An End-to-End Discriminative Approach to Machine Translation.

Word alignment and SMT performance

Alexander Fraser and Daniel Marcu. Semi-Supervised Training for Statistical Word Alignment.

- The AER metric is broken. Use F-measure instead
- AER is not a good indicator of MT performance
- F-measure is a good indicator of MT performance
- Discriminatively train parameters to maximize unbalanced F-measure

Decoding

Taro Watanabe, Hajime Tsukada and Hideki Isozaki. Left-to-Right Target Generation for Hierarchical Phrase-Based Translation.

Syntax-based SMT

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang and Ignacio Thayer. Scalable Inference and Training of Context-Rich Syntactic Translation Models.

- Bigger rules are better than minimal rules.
- Incorporating phrases

Word clustering

David Talbot and Miles Osborne. Modelling Lexical Redundancy for Machine Translation.

Clustering of source words

Word alignment

Benjamin Wellington, Sonjia Waxmonsky and I. Dan Melamed. Empirical Lower Bounds on the Complexity of Translational Equivalence

- Translation equivalence without gap is relatively small
- Using discontinuous constituents leads to higher coverage
- Many sentendce pairs cannot be hierarchically aligned without discontinuities

EBMT incorporating SMT techniques

- End-to-End discriminative approach

  The key technique is representing a candidate translation as a feature vector. cf. (Liang et.al 2006)

  perceptron, structured SVM, ... Searn might be good if we can formulate EBMT as a sequence of classification

- Word, phrase, or dependency alignment directly connected to decoding.

  For example, use alignments that are used in N-best translation candidates in discriminative training.

  cf. (Alexander Fraser 2006, Nagata et.al 2006)